Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI

Year: 2021 | Citations: 363 | Authors: J. M. Durán, K. Jongsma

Abstract

The use of black box algorithms in medicine has raised scholarly concerns due to their opaqueness and lack of trustworthiness. Concerns about potential bias, accountability and responsibility, patient autonomy and compromised trust transpire with black box algorithms. These worries connect epistemic concerns with normative issues. In this paper, we outline that black box algorithms are less problematic for epistemic reasons than many scholars seem to believe. By outlining that more transparency in algorithms is not always necessary, and by explaining that computational processes are indeed methodologically opaque to humans, we argue that the reliability of algorithms provides reasons for trusting the outcomes of medical artificial intelligence (AI). To this end, we explain how computational reliabilism, which does not require transparency and supports the reliability of algorithms, justifies the belief that results of medical AI are to be trusted. We also argue that several ethical concerns remain with black box algorithms, even when the results are trustworthy. Having justified knowledge from reliable indicators is, therefore, necessary but not sufficient for normatively justifying physicians to act. This means that deliberation about the results of reliable algorithms is required to find out what is a desirable action. Thus understood, we argue that such challenges should not dismiss the use of black box algorithms altogether but should inform the way in which these algorithms are designed and implemented. When physicians are trained to acquire the necessary skills and expertise, and collaborate with medical informatics and data scientists, black box algorithms can contribute to improving medical care.